Psychonomic Bulletin & Review
2008, 15 (5), 1031-1034
doi: 10.3758/PBR.15.5.1031

Decision noise may mask criterion shifts: Reply to Balakrishnan and MacDonald (2008)

Christoph T. Weidemann
University of Pennsylvania, Philadelphia, Pennsylvania

and

Shane T. Mueller
Klein Associates Division,
Applied Research Associates, Fairborn, Ohio

J. D. Balakrishnan and J. A. MacDonald (2008) argue that RT-based measures of signal detection processes provide evidence against signal detection theory's notion of a flexible decision criterion. They argue that this evidence is immune to the alternative explanation proposed by S. T. Mueller and C. T. Weidemann (2008), that decision noise may mask criterion shifts. We show that noise in response times can produce the same effects as are produced by noise in confidence ratings. Given these results, the evidence is not sufficient to categorically reject the notion of a flexible response policy implemented through shifts in a decision criterion.

Theories of signal detection attempt to distinguish perceptual factors from decision effects (such as response biases). To that effect, signal detection theory (SDT) incorporates the notion of a flexible decision criterion that adapts to contingencies in the environment (such as stimulus base rates or pay-off schemes). Balakrishnan (1998a, 1998b, 1999) identified serious problems with SDT that render derived measures of sensitivity and bias (such as d' and β) suspect. On the basis of novel measures of bias, Balakrishnan (1998a, 1998b, 1999) argued that the notion of such a flexible decision criterion that shifts in response to stimulus contingencies is fundamentally flawed.

We have shown that Balakrishnan's (1998a) results, as well as those from other experiments, can in fact be accounted for by a simple model incorporating a flexible decision criterion (Mueller & Weidemann, 2008). This model is an extension of SDT, but in addition to perceptual noise, it includes noise in the mapping from percepts to responses (*decision noise*). In particular, we distinguish between two types of decision noise: *classification noise* and *confidence noise*. *Classification noise* refers to noise in the mapping between the percept and the classification response (e.g., *yes–no*, *present–absent*, *A–B*), whereas *confidence noise* refers to the noise in the mapping between an internal state and a dependent measure indexing response confidence (such as confidence ratings). We view the decision noise model (DNM) as constituting a simple existence proof (rather than a full-fledged alternative to existing theories) that the results from signal detection tasks[1] are compatible with flexible decision criteria that adapt to task contingencies. In the DNM, confidence noise can be larger than classification noise, which is consistent with the results of our receiver operating characteristic (ROC) analyses comparing response-related distal-stimulus ROC functions and confidence ROC functions (see Mueller & Weidemann, 2008, for details). As we demonstrated earlier (Mueller & Weidemann, 2008), this discrepancy between classification noise and confidence noise can mask shifts in decision criteria.

Balakrishnan and MacDonald (2008) criticize our work on the basis of three major points. Specifically, they argue that:

1. Our account of how decision noise may mask criterion shifts should not apply to cases where classification responses are collected without confidence ratings, but measures based on response times (RTs) of yes–no responses do show patterns similar to those for corresponding measures based on confidence ratings.
2. The DNM is underconstrained because it can predict arbitrary likelihood ratios between the two classification responses, whereas empirical likelihood ratios at that point seem to be constrained to equal and smoothly approach 1.0.
3. Conditioning the analyses on the previous classification response should reduce decision noise, but fails to show consistent increases in d'.

These criticisms center on details of dependent variables (specifically RT and sequential dependencies) that we did not attempt to model in our previous article (Mueller & Weidemann, 2008) and on questions about the parsimony of our model (i.e., is it constrained to produce a certain result, rather than just able to). In what follows, we address each of these criticisms in detail. To foreshadow our main points, we argue that:

1. Noise in the mapping between an internal state and any dependent measure qualifying the classification response (including RTs) can produce the same effects that we have demonstrated for confidence ratings.
2. Empirical likelihood ratios between the classification responses need not smoothly approach 1.0 and can even deviate from 1.0 in some cases. The DNM can produce likelihood ratios that smoothly approach 1.0, and its lack of constraint to always do so is warranted by the data.
3. Conditioning analyses on the previous classification responses produces extremely small ($<.06$) changes in d', and the direction of change even within a base rate condition can differ for different response instructions (yes–no vs. confidence ratings). We explicitly did not attempt to model sequential dependencies and did not specify any relationship between the previous response and decision criteria.

We therefore maintain that criterion shifts may be masked by noise in the mapping between internal states and responses, even when no confidence ratings are collected.

C. T. Weidemann, ctw@cogsci.info

Decision Noise and Response Time

Using response time to qualify the classification response, Balakrishnan and MacDonald analyzed those conditions in our experiment (Mueller & Weidemann, 2008) that required only a classification response (but no confidence ratings). Their results replicated those from our analyses that used confidence ratings to qualify the classification response. Balakrishnan and MacDonald argued that, unlike the confidence rating results, the RT results could not be explained by differential levels of decision noise, because only one type of response was required.

Crucial to our account of confidence rating data (Mueller & Weidemann, 2008) was the notion of different amounts of decision noise for classification responses and dependent measures indexing response confidence. Confidence ratings are not the only conceivable measure of response confidence, and RTs in particular could reasonably be used for this purpose (with faster RTs often indexing more confident responses). Confidence noise, therefore, is not specific to confidence ratings but naturally extends to other dependent variables.

The notion of response confidence, however, does not need to be invoked. Many dependent measures (e.g., RTs, or electrophysiological or brain imaging data) could qualify a classification response and could reasonably be used to generate ROC or likelihood ratio functions. To the extent that these measures are noisy indices of internal states, they could produce patterns of results that are similar to those observed for confidence ratings. Depending on the nature of these measures, "confidence noise" or even "decision noise" may not be the most appropriate terms to describe the noise associated with them, but nevertheless their effects could mirror those of decision noise.

Indeed, even when RTs are generated randomly, the main RT results (i.e., crossing RT-ROC functions and likelihood functions that pass through 1.0 at the middle RT bin regardless of base rate) can be captured simply by setting the hit and false alarm rates as well as the stimulus base rates to the actual values. These assumptions are consistent with a shift in a decision criterion (although they do not necessarily imply it), and they illustrate that the main aspects of the data can be captured even with very basic assumptions. As we show below, with the additional constraint (confirmed in our data) that RTs for correct responses are slightly less variable than those for incorrect responses, one can capture the more subtle features of the RT likelihood ratio function—namely, the smooth transition between values below and above 1.0, and the small tendency to approach 1.0 for extreme (fast or slow) RTs.

To illustrate how decision noise can produce RT-ROC functions that are similar to the confidence ROC (C-ROC) functions that we observed (Mueller & Weidemann, 2008, Figure 9) and to RT-likelihood ratio functions like those reported by Balakrishnan and MacDonald (Figure 2), we simulated data from a classification experiment with hit and false alarm rates equivalent to those in our experiment. We randomly generated RTs for all conditions by sampling from normal distributions with means of 1,000 msec and standard deviations of 100 msec for correct trials and 110 msec for incorrect trials. These assumptions lead to relatively more incorrect responses in the tails of the distributions, which is necessary to capture the smooth transition between values below and above 1.0 and the small tendency for the likelihood ratios to approach 1.0 for extreme (fast or slow) RTs. To validate the assumptions for this simulation, we confirmed that our data indeed showed higher variability for incorrect trials and that the distribution of RTs was approximately log normal. For each condition, we simulated 200,000 trials with "yes" and "no" response proportions set equal to the empirical hit and false alarm rates and stimulus base rates in both 2:1 and 1:2 proportions. We then recoded and binned the simulated RTs as Balakrishnan and MacDonald did, in order to obtain the RT-ROC and RT-likelihood ratio functions shown in Figures 1 and 2 (we used a bin size of 5,000 samples when the entire simulated data set was sorted by RT, which resulted in 2.5% of the data falling into any RT bin).

We note that the qualitative pattern of the simulated RT-ROC functions in Figure 1 is very similar to that seen in the confidence data (Mueller & Weidemann, 2008, Figure 9). Likewise, the simulated RT likelihood ratio functions (Figure 2) are consistently below 1.0 for "A" responses and consistently above 1.0 for "B" responses irrespective of stimulus base rates, much like the actual data (Balakrishnan & MacDonald, Figure 2). We made the assumptions described above primarily for simplicity and convenience. The fact that these patterns can be observed for randomly generated RTs with a slight difference in variance for correct and incorrect trials shows that they do not depend on sophisticated assumptions about the mapping between internal states and RTs. It is also obvious that the simple model described above cannot account for every detail of the empirical likelihood ratio functions shown by Balakrishnan

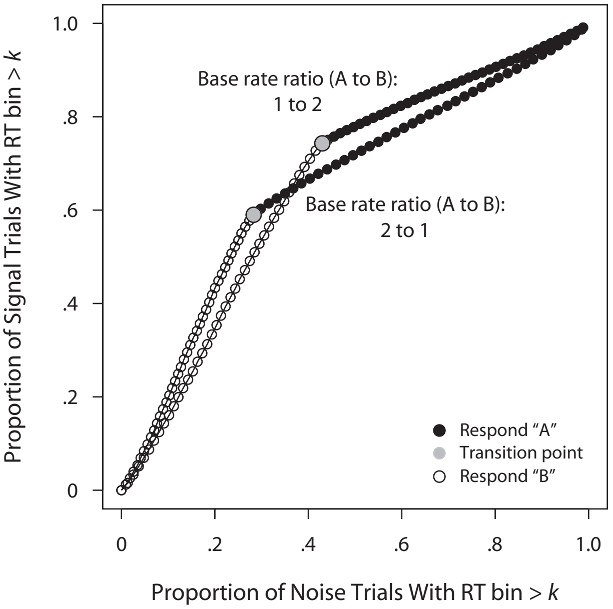

Figure 1. Simulated response time (RT) receiver operating characteristic (ROC) functions. These simulated functions show the same qualitative pattern as those for the ROC functions based on confidence ratings (Mueller & Weidemann, 2008).

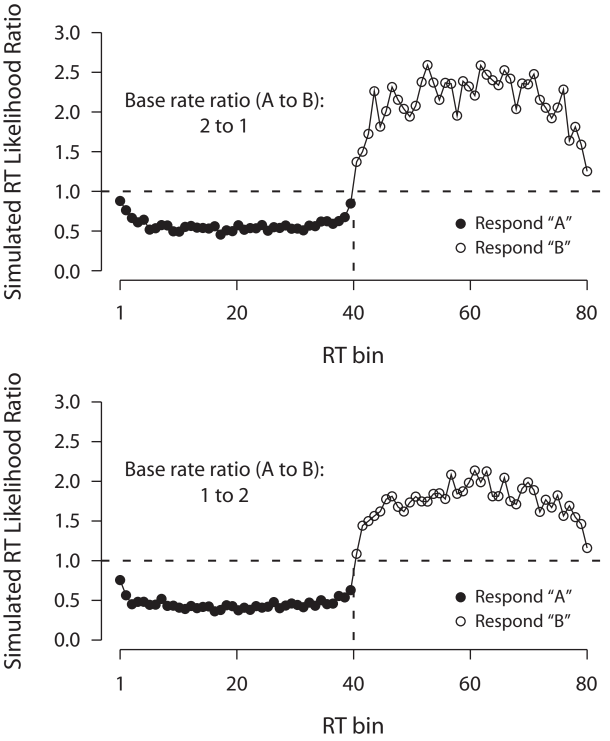

Figure 2. Simulated response time (RT) likelihood ratio functions. As in the RT data (Balakrishnan & MacDonald, 2008), the likelihood ratio is consistently below 1 for "A" responses and consistently above 1 for "B" responses, irrespective of stimulus base rate.

and MacDonald (Figure 2). Particularly the empirical functions seem somewhat noisier and also seem to approach 1.0 equally from both sides at the transition between classification responses, whereas the tendency to approach 1.0 at the transition is more pronounced for the more frequent classification responses in our simulations.

These simulations suggest that more realistic models of RTs in signal detection tasks do not necessarily have to shun the notion of flexible decision criteria to account for RT likelihood ratio functions. The assumption that RTs for correct responses are less variable than those for incorrect responses is admittedly ad hoc (although confirmed by our data); it serves only to smooth the transition in the likelihood ratios and could be implemented in process models with a number of reasonable mechanisms.

Constraints in the Likelihood Ratio Function

Classical SDT predicts that the likelihood ratio should equal 1.0 at the point between the two classification responses for unbiased responses and that it should deviate from 1.0 if the response is biased. The likelihood ratio at any point is equivalent to the slope of the ROC function at that point (see Zhang & Mueller, 2005), and therefore any peak in the ROC function where the slope changes from below 1.0 to above 1.0 (see Mueller & Weidemann, 2008, Figures 9 and 12) corresponds to the point where the likelihood ratio crosses 1.0. Balakrishnan and MacDonald (2008) argue that this point is rarely, if ever, found to deviate from the point between the classification responses (such a deviation would indicate a biased decision rule, according to classical SDT) and furthermore argue that the likelihood ratios associated with low confidence responses tend to be very close to 1.0. As Balakrishnan and MacDonald point out, the DNM can produce these results, but it is not constrained to do so—a point that they view as a disadvantage of the DNM.

Whereas the results described above (likelihood ratios transitioning from below to above 1.0 between the response categories and approaching 1.0 for low confidence ratings) have been found repeatedly, they are by no means universal. Figure 3 shows the ratios of the smallest confidence likelihood ratio above 1.0 and the largest below 1.0 in the data presented by Van Zandt (2000) and modeled with the DNM by us (Mueller & Weidemann, 2008). For example, a ratio of 4.0 indicates that the likelihood ratio changed by a factor of 4.0 (e.g., from 0.5 to 2.0) between the confidence responses flanking the neutral likelihood ratio (usually the two lowest confidence ratings for either classification response). Most of these ratios are close to 1.0, but several ratios on the upper end of the scale shown in Figure 3 indicate that the neutral likelihood ratio is not always smoothly approached. In light of these data, a model constrained in the ways suggested by Balakrishnan and MacDonald does not seem warranted. Indeed, our model predicts that violations of these constraints should be observable to the extent that the response policy can be shifted substantially, especially if the difference between classification and confidence noise can be reduced.

Sequential Dependencies

We identified substantial sequential dependencies in our data, which may represent one of presumably many sources of decision noise (Mueller & Weidemann, 2008). More specifically, we showed that participants in our experiment were likely to repeat the previous confidence rating even when the classification response changed (e.g., a high-confidence "A" response was more likely to be followed by a high-confidence "B" response than by a low-confidence "B" response). Because the source of the decision noise was not crucial for our argument, the DNM does not incorporate any sequential dependencies and we specifically did not attempt to account for any data by assuming that a classification criterion shifts back and forth on each trial depending on the previous response, as suggested by Balakrishnan and MacDonald.

Balakrishnan and MacDonald calculated d' for trials conditioned on the prior classification response and found small ($<.06$) differences within each base rate and response (confidence rating vs. forced choice) condition. In particular, the overall d' values fell between the d' values obtained for "A" responses and for "B" responses, although the relative order of d' values was not consistent across base rate or response conditions. Balakrishnan and MacDonald argue that, contrary to the data, our model should predict consistently higher d' values when conditioning on the previous response, because this should reduce decision noise.

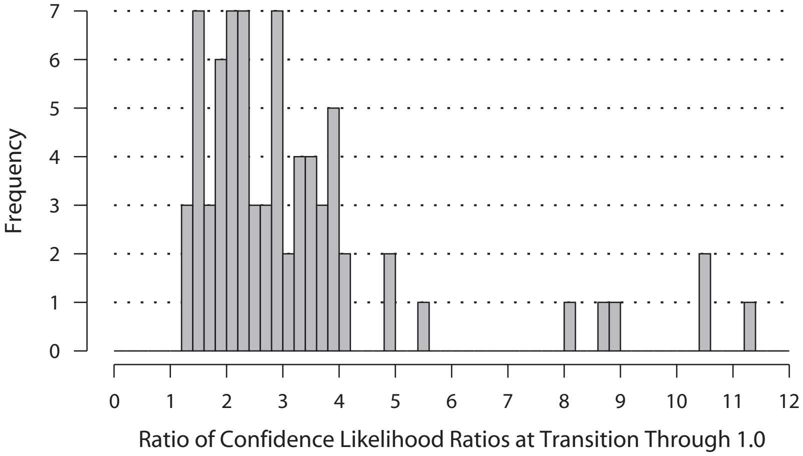

Figure 3. Ratios of the smallest confidence likelihood ratio greater than 1 and the largest confidence likelihood ratio less than 1 for the data reported in Van Zandt (2000) and modeled with the decision noise model in Mueller and Weidemann (2008). The transition through 1.0 usually occurred (with some exceptions) between the two response classes. The ratio is between adjacent likelihood ratios except for rare cases when a likelihood ratio was exactly 1.0 (in such cases the ratio was taken between the two values flanking the 1.0 likelihood ratio).

To the extent that these small differences in d' values are reliable, their relative orders would indeed be difficult to model, because it seems to interact with base rate and response conditions. In particular, as Balakrishnan and MacDonald (2008) pointed out, these data seem at odds with a simple model that adjusts a decision criterion up or down on every trial depending on the previous classification response. We do not have an account for these results, but we note that more complex effects of the previous response on a decision criterion (possibly contingent on whether or not the previous response was correct) may be able to explain the small fluctuations in hit and false alarm rates that give rise to these results.

Discussion

As we have shown above and previously (Mueller & Weidemann, 2008), the indices of criterion shifts proposed by Balakrishnan (1998a, 1998b, 1999) are not always able to detect such shifts in the presence of decision noise. We have shown that decision noise does indeed seem to have a large influence on confidence rating and forced choice responses (Mueller & Weidemann, 2008) and have argued that the RT results presented by Balakrishnan and MacDonald (2008) are consistent with a high degree of noise in the mapping between internal states and RT for a two-alternative forced choice response.

As we stated previously (Mueller & Weidemann, 2008), we do not mean to argue in favor of wholesale acceptance of SDT or against sequential sampling models. With the DNM, we simply showed that the data are compatible with flexible decision criteria that adapt to task contingencies (Mueller & Weidemann, 2008). To understand the processes involved in signal detection better, it is crucial to carefully analyze the extent and the limits of the implications of theory violations. The outcomes of such analyses provide important guidance and constraints for more adequate models of choice under uncertainty.

AUTHOR NOTE

This research was supported in part by a postdoctoral fellowship to C.T.W. from the German Academic Exchange Service (DAAD). Correspondence may be sent via e-mail to C. T. Weidemann, Department of Psychology, University of Pennsylvania (ctw@cogsci.info).

REFERENCES

Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, **40**, 601-623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, **3**, 68-90.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception & Performance*, **25**, 1189-1206.

Balakrishnan, J. D., & MacDonald, J. A. (2008). Decision criteria do not shift: Commentary on Mueller and Weidemann (2008a). *Psychonomic Bulletin & Review*, **15**, 1022-1030.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, **15**, 465-494.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 582-600.

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and nonparametric estimate of sensitivity. *Psychometrika*, **70**, 203-212.

NOTE

[1] We use this term in the loose sense defined earlier (Mueller & Weidemann, 2008, note 1) to refer to the basic paradigm in which two stimulus classes are discriminated and categorized into two classes (old or new, signal or noise, yes or no, A or B, etc.).

(Manuscript received May 5, 2008;
revision accepted for publication July 19, 2008.)